第26卷 第5期
2009年10月

工 程 数 学 学 报
CHINESE JOURNAL OF ENGINEERING MATHEMATICS

Vol. 26 No. 5
Oct. 2009

文章编号:1005-3085(2009)05-0861-05

双向分类随机效应模型中方差分量的估计*

许王莉

(中国人民大学统计学院，北京 100872)

摘 要: 双向分类随机效应模型是一类有着广泛应用背景的统计模型，其中对模型中方差参数的一种重要估计方法是方差分析估计。由于方差分析法得到估计的均方误差 (MSE) 并不是最小的，本文在一类新的估计族中提出了改进的 ANOVA 估计。结果表明新的估计比 ANOVA 具有较小的 MSE，这种新的估计方法可推广到医学领域中常见的一般模型。

关键词: 双向分类随机效应模型；ANOVA 估计；MSE

分类号: AMS(2000) 62J05　　**中图分类号**: O212.1　　**文献标识码**: A

1 引言

随机效应模型是广泛应用于生存分析、医学、经济、质量控制等领域的一类模型，对模型中方差参数估计的研究已经很多，比如方差分析估计 (ANOVA)，极大似然估计 (MLE)，限制极大似然估计 (REMLE)，最小范数二次无偏估计 (MINQUE)、谱分解估计 (SDE)，Bayes 估计等。由于 ANOVA 估计具有显示表达式，关于它的研究文献也比较多，比如文献 [1,2]。

双向分类随机效应模型是一类重要的随机效应模型，带交互效应的双向分类随机效应模型的具体形式为

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \quad i = 1, \cdots, a; \quad j = 1, \cdots, b; \quad k = 1, \cdots, c, \tag{1}$$

这里 $a \geq 2$, $b \geq 2$, $c \geq 2$, μ 是未知的固定常数，α_i, β_j, γ_{ij} 分别表示固定效应，随机效应和交叉效应，e_{ijk} 是随机误差。假定 β_j, γ_{ij}, e_{ijk} 为相互独立的正态随机变量，其均值均为零，方差分别为 σ_β^2, σ_γ^2, σ_e^2, y_{ijk} 为第 (i,j) 单元上的第 k 次观测值。

模型 (1) 的矩阵形式为

$$y = \mathbf{1}_n \mu + X\alpha + U_1\beta + U_2\gamma + e, \tag{2}$$

其中

$$y = (y_1^\tau, y_2^\tau, \cdots, y_a^\tau)^\tau, \quad y_i = (y_{i11}, y_{i12}, \cdots, y_{i1c}, y_{i21}, \cdots, y_{i2c}, \cdots, y_{ib1}, \cdots, y_{ibc})^\tau,$$

$$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_a), \quad \beta = (\beta_1, \beta_2, \cdots, \beta_b), \quad \gamma = (\gamma_{11}, \gamma_{12}, \cdots, \gamma_{1b}, \gamma_{21}, \cdots, \gamma_{ab}),$$

$$e = (e_1^\tau, e_2^\tau, \cdots, e_a^\tau)^\tau, \quad e_i = (e_{i11}, e_{i12}, \cdots, e_{i1c}, e_{i21}, \cdots, e_{i2c}, \cdots, e_{ib1}, \cdots, e_{ibc})^\tau.$$

收稿日期: 2007-08-10.　**作者简介**: 许王莉 (1978年7月生)，女，博士，讲师. 研究方向：纵向数据分析.

***基金项目**: 国家自然科学基金 (10701079)；教育部人文社会科学研究项目 (08JC910002).

$X = I_a \otimes \mathbf{1}_{ab}$，$U_1 = \mathbf{1}_a \otimes I_b \otimes \mathbf{1}_c$，$U_2 = I_{ab} \otimes \mathbf{1}_c$，$\otimes$表示Keonecker乘积，$\mathbf{1}_m$和$I_m$分别表示元素全为1的$m$维列向量和$m$维单位矩阵。根据模型(1)的假定，不难得出$\beta, \gamma, e$相互独立且$\beta \sim N(0,\sigma_\beta^2 I_b)$，$\gamma \sim N(0,\sigma_\gamma^2 I_{ab})$，$e \sim N(0,\sigma_e^2 I_n)$，这里$n = abc$。

对双向分类随机效应模型(2)，方差分量σ_β^2，σ_γ^2，σ_e^2的ANOVA估计分别记为$\hat{\sigma}_{\beta A}^2$，$\hat{\sigma}_{\gamma A}^2$和$\hat{\sigma}_{eA}^2$，它们是二次统计量的线性组合。范和王[3]在均方损失下修正了ANOVA估计，得到的估计无妨记为$\hat{\sigma}_{\beta M_1}^2$，$\hat{\sigma}_{\gamma M_1}^2$和$\hat{\sigma}_{eM_1}^2$，它们仍然是二次统计量的线性组合。事实上，范和王[3]在估计类$\{t\hat{\sigma}_{\beta A}^2 | t > 0\}$，$\{t\hat{\sigma}_{\gamma A}^2 | t > 0\}$，$\{t\hat{\sigma}_{eA}^2 | t > 0\}$中分别考虑了在均方损失下优于ANOVA的估计。受到Kelly和Mathew[4]关于在混合效应模型中构造非负方差参数估计方法的启发，本文在一类新的估计族中构造方差参数σ_β^2，σ_γ^2的修正ANOVA估计，得到的估计分别记为$\hat{\sigma}_{\beta M_2}^2$，$\hat{\sigma}_{\gamma M_2}^2$。在均方损失意义下，$\hat{\sigma}_{\beta M_2}^2$和$\hat{\sigma}_{\gamma M_2}^2$分别一致优于范和王[3]的估计$\hat{\sigma}_{\beta M_1}^2$和$\hat{\sigma}_{\gamma M_1}^2$，以及ANOVA估计$\hat{\sigma}_{\beta A}^2$和$\hat{\sigma}_{\gamma A}^2$。

为了与范和王[3]提出的修正ANOVA估计做比较，本文第2小节关于方差分量的ANOVA估计都是建立在范和王[3]文章中所研究的双向分类随机效应模型上，文章第3把结论推广到更一般的双向分类随机效应模型。

2 方差分量 ANOVA 估计的改进

为了构造改进的ANOVA估计，首先引入ANOVA估计σ_β^2，σ_γ^2，σ_e^2。定义统计量

$$SS_\beta = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}),$$

$$SS_\gamma = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}),$$

$$SS_e = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(y_{ijk} - \bar{y}_{ij\cdot}),$$

其中

$$\bar{y}_{ij\cdot} = \sum_{k=1}^{c} y_{ijk}/c, \quad \bar{y}_{i\cdot\cdot} = \sum_{j=1}^{b}\sum_{k=1}^{c} y_{ijk}/bc, \quad \bar{y}_{\cdot j\cdot} = \sum_{i=1}^{a}\sum_{k=1}^{c} y_{ijk}/ac, \quad \bar{y}_{\cdots} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} y_{ijk}/abc.$$

根据范和王[3]，SS_β，SS_γ，SS_e可以分别表示为$y^\tau H_\beta^\tau H_\beta y$，$y^\tau H_\gamma^\tau H_\gamma y$和$y^\tau H_e^\tau H_e y$。记$z_\beta = H_\beta y$，$z_\gamma = H_\gamma y$和$z_e = H_e y$，则有

1) z_β，z_γ，z_e相互独立;

2) $z_\beta^\tau z_\beta \sim \lambda_\beta^2 \chi_{r_\beta}^2$，$z_\gamma^\tau z_\gamma \sim \lambda_\gamma^2 \chi_{r_\gamma}^2$，$z_e^\tau z_e \sim \lambda_e^2 \chi_{r_e}^2$，这里

$$r_\beta = b-1, \quad r_\gamma = (a-1)(b-1), \quad r_e = ab(c-1),$$

$$\lambda_\beta^2 = ac\sigma_\beta^2 + c\sigma_\gamma^2 + \sigma_e^2, \quad \lambda_\gamma^2 = c\sigma_\gamma^2 + \sigma_e^2, \quad \lambda_e^2 = \sigma_e^2.$$

令SS_β，SS_γ，SS_e和它们的数学期望相等，就得到一组关于σ_β^2，σ_γ^2，σ_e^2的方程组，解此方程组得方差分量的ANOVA估计分别为

$$\hat{\sigma}_{\beta A}^2 = \frac{SS_\beta}{acr_\beta} - \frac{SS_\gamma}{acr_\gamma}, \quad \hat{\sigma}_{\gamma A}^2 = \frac{SS_\gamma}{cr_\gamma} - \frac{SS_e}{cr_e}, \quad \hat{\sigma}_{eA}^2 = \frac{SS_e}{r_e}.$$

范和王[3] 在估计类 $\{t\hat{\sigma}^2_{eA}\,|\,t>0\}$，$\{t\hat{\sigma}^2_{\beta A}\,|\,t>0\}$，$\{t\hat{\sigma}^2_{\gamma A}\,|\,t>0\}$ 中分别考虑了在均方损失下优于 ANOVA 的估计 $\hat{\sigma}^2_{eA}$，$\hat{\sigma}^2_{\beta A}$，$\hat{\sigma}^2_{\gamma A}$，得到的估计 $\hat{\sigma}^2_{\beta M_1}$，$\hat{\sigma}^2_{\gamma M_1}$，$\hat{\sigma}^2_{eM_1}$ 具有下述结论：

在均方误差意义下，σ^2_β 的估计

$$\hat{\sigma}^2_{\beta M_1}=\big(r_\beta/(r_\beta+2)\big)\hat{\sigma}^2_{\beta A},$$

σ^2_γ 的估计

$$\hat{\sigma}^2_{\gamma M_1}=\big(r_\gamma/(r_\gamma+2)\big)\hat{\sigma}^2_{\gamma A},$$

σ^2_e 的估计

$$\hat{\sigma}^2_{eM_1}=\big(1/(r_e+2)\big)\hat{\sigma}^2_{eA},$$

分别优于 ANOVA 估计 $\hat{\sigma}^2_{\beta A}$，$\hat{\sigma}^2_{\gamma A}$ 和 $\hat{\sigma}^2_{eA}$。

对 σ^2_β 和 σ^2_γ 的估计，受到 Kelly 和 Mathew[4] 关于在混合效应模型中构造非负方差参数估计方法的启发，本文提出一个新的改进的 ANOVA 估计，这个估计在均方损失下一致优于相应的范和王[3] 提出的修正 ANOVA 估计和 ANOVA 估计。关于 σ^2_β 和 σ^2_γ，分别考虑具有下述形式的估计类

$$\mathcal{L}_\beta=\Big\{\frac{m_{\beta1}}{ac}\Big(\frac{z^\tau_\beta z_\beta}{r_\beta}-m_{\beta2}\frac{z^\tau_\gamma z_\gamma}{r_\gamma}\Big)\,\Big|\,0<m_{\beta1}<1,\ m_{\beta2}<1\Big\},$$

$$\mathcal{L}_\gamma=\Big\{\frac{m_{\gamma1}}{c}\Big(\frac{z^\tau_\gamma z_\gamma}{r_\gamma}-m_{\gamma2}\frac{z^\tau_e z_e}{r_e}\Big)\,\Big|\,0<m_{\gamma1}<1,\ m_{\gamma2}<1\Big\}.$$

事实上，在上述估计类 $\mathcal{L}_\beta$, $\mathcal{L}_\gamma$ 中，$m_{\beta2}<1$, $m_{\gamma2}<1$ 使得在这两个估计类中得到的估计，无妨记为 $\hat{\sigma}^2_{\beta M_2}$ 和 $\hat{\sigma}^2_{\gamma M_2}$，它们取负值的概率分别小于 $\hat{\sigma}^2_{\beta M_1}$ 和 $\hat{\sigma}^2_{\gamma M_1}$，以及 $\hat{\sigma}^2_{\beta A}$ 和 $\hat{\sigma}^2_{\gamma A}$。

首先考虑 σ^2_β 的估计，研究在估计类 $\mathcal{L}_\beta$ 中选择合适 $m_{\beta1}$, $m_{\beta2}$ 的方法。实际上，本文通过最小化估计类 $\mathcal{L}_\beta$ 的均方误差 (MSE) 的 σ^4_β 和 λ^4_γ 的系数得到 $m_{\beta1}$, $m_{\beta2}$。首先计算 MSE$(\mathcal{L}_\beta)$，简单推导得

$$\begin{aligned}\mathrm{MSE}(\mathcal{L}_\beta)&=\mathrm{Var}(\mathcal{L}_\beta)+\big(E(\mathcal{L}_\beta)-\sigma^2_\beta\big)^2\\&=\Big\{\frac{2m^2_{\beta1}}{r_\beta}+(m_{\beta1}-1)^2\Big\}\sigma^4_\beta+\frac{m^2_{\beta1}}{(ac)^2}\Big\{\frac{2}{r_\beta}+\frac{2m^2_{\beta2}}{r_\gamma}+(1-m_{\beta2})^2\Big\}\lambda^4_\gamma\\&\quad+\frac{2m_{\beta1}}{ac}\Big\{\frac{2m_{\beta1}}{r_\beta}+(m_{\beta1}-1)(1-m_{\beta2})\Big\}\lambda^2_\gamma\sigma^2_\beta.\end{aligned}$$

不难证明，最小化 σ^4_β 的系数，以及给定 $m_{\beta1}$ 最小化 λ^4_γ 的系数，可得 $m_{\beta1}=r_\beta/(r_\beta+2)$ 和 $m_{\beta2}=r_\gamma/(r_\gamma+2)$。最终得到 σ^2_β 的估计为

$$\hat{\sigma}^2_{\beta M_2}=\frac{m_{\beta1}}{ac}\Big(\frac{z^\tau_\beta z_\beta}{r_\beta}-m_{\beta2}\frac{z^\tau_\gamma z_\gamma}{r_\gamma}\Big)=\frac{r_\beta}{ac(r_\beta+2)}\Big(\frac{z^\tau_\beta z_\beta}{r_\beta}-\frac{r_\gamma}{r_\gamma+2}\times\frac{z^\tau_\gamma z_\gamma}{r_\gamma}\Big).$$

为了符号上的方便，分别用 $g_1(m_{\beta1})$，$g_2(m_{\beta1},m_{\beta2})$ 和 $g_{12}(m_{\beta1},m_{\beta2})$ 表示 MSE$(\mathcal{L}_\beta)$ 中 σ^4_β，λ^4_γ 和 $\lambda^2_\gamma\sigma^2_\beta$ 的系数。对于估计 $\hat{\sigma}^2_{\beta M_2}$，有如下定理。

定理 1 在均方损失意义下，估计 $\hat{\sigma}^2_{\beta M_2}$ 优于估计 $\hat{\sigma}^2_{\beta A}$ 和 $\hat{\sigma}^2_{\beta M_1}$。

证明 先证估计 $\hat{\sigma}^2_{\beta M_2}$ 优于 ANOVA 估计 $\hat{\sigma}^2_{\beta A}$。对于 ANOVA 估计，$m_{\beta1}=1$, $m_{\beta2}=1$。对估计 $\hat{\sigma}^2_{\beta M_2}$，$m_{\beta1}=r_\beta/(r_\beta+2)$, $m_{\beta2}=r_\gamma/(r_\gamma+2)$。根据上述分析，

$$g_1(1)>g_1(r_\beta/(r_\beta+2)),\quad g_2(1,1)>g_2(r_\beta/(r_\beta+2),r_\gamma/(r_\gamma+2)).$$

经过简单推导可得 $g_{12}(1,1) = 4/(acr_\beta) > g_{12}(r_\beta/(r_\beta+2), r_\gamma/(r_\gamma+2))$。故结论成立。

接下来说明 $\hat{\sigma}^2_{\beta M_2}$ 优于估计 $\hat{\sigma}^2_{\beta M_1}$。对估计 $\hat{\sigma}^2_{\beta M_1}$，$m_{\beta 1} = r_\beta/(r_\beta+2)$，$m_{\beta 2} = 1$。对于这两个估计，$\sigma^4_\beta$ 的系数 $g_1(m_{\beta 1})$ 相同，我们说明

$$g_2(r_\beta/(r_\beta+2), 1) > g_2(r_\beta/(r_\beta+2), r_\gamma/(r_\gamma+2)),$$

$$g_{12}(r_\beta/(r_\beta+2), 1) = 4/(acr_\beta) > g_{12}(r_\beta/(r_\beta+2), r_\gamma/(r_\gamma+2)).$$

简单推导可得

$$\begin{aligned} g_2(r_\beta/(r_\beta+2), r_\gamma/(r_\gamma+2)) &= \frac{r_\beta^2}{(ac)^2(r_\beta+2)^2}\left\{\frac{2}{r_\beta} + \frac{2}{r_\gamma}\frac{r_\gamma^2+2r^2}{(r_\gamma+2)^2}\right\} \\ &< \frac{r_\beta^2}{(ac)^2(r_\beta+2)^2}\left\{\frac{2}{r_\beta} + \frac{2}{r_\gamma}\right\} = g_2\big(r_\beta/(r_\beta+2), 1\big), \end{aligned}$$

类似推导可得 $g_{12}(r_\beta/(r_\beta+2), r_\gamma/(r_\gamma+2)) < g_{12}(r_\beta/(r_\beta+2), 1)$。

关于 σ^2_γ，在估计类 $\mathcal{L}_\gamma$ 中，ANOVA 估计 $\hat{\sigma}^2_{\gamma A}$ 对应于 $m_{\gamma 1} = 1$，$m_{\gamma 2} = 1$ 的情况。范和王[3] 的修正 ANOVA 估计对应于

$$m_{\gamma 1} = r_\gamma/(r_\gamma+2), \quad m_{\gamma 2} = 1.$$

类似于估计 $\hat{\sigma}^2_{\beta M_2}$ 中选择 $m_{\beta 1}$，$m_{\beta 2}$ 的方法，经过推导可得 $m_{\gamma 1} = r_\gamma/(r_\gamma+2)$，$m_{\gamma 2} = r_e/(r_e+2)$，也就是，我们得到新的 σ^2_γ 估计为

$$\hat{\sigma}^2_{\gamma M_2} = \frac{r_\gamma}{c(r_\gamma+2)}\Big(\frac{z_\gamma^\tau z_\gamma}{r_\gamma} - \frac{r_e}{r_e+2} \times \frac{z_e^\tau z_e}{r_e}\Big).$$

对估计 $\hat{\sigma}^2_{\gamma M_2}$，有下述定理。

定理 2 在均方损失意义下，估计 $\hat{\sigma}^2_{\gamma M_2}$ 优于估计 $\hat{\sigma}^2_{\gamma A}$ 和 $\hat{\sigma}^2_{\gamma M_1}$。

类似定理 1 的证明过程，不难得到定理 2 的结论，具体细节就不再累述。

3 一般双向分类随机效应模型中 ANOVA 估计的改进

第 1，2 小节分析了对双向分类随机效应模型 (2) 的方差分量的估计。模型中假定 α 为固定效应，在很多情况下 α 假定为随机效应，也就是假定 $\alpha \sim N(\sigma_3^2, I_a)$ 且与 β，γ，e 独立。此时如何给出方差分量的修正 ANOVA 估计，是本节主要讨论的问题。

在给出改进的 ANOVA 估计之前，先引入模型的 ANOVA 估计。定义统计量

$$SS_\alpha = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot}),$$

这里

$$\bar{y}_{i\cdot\cdot} = \sum_{j=1}^b \sum_{k=1}^c y_{ijk}/bc.$$

由文献 [5] §10.2，SS_α 可分解为 $y^\tau H_\alpha^\tau H_\alpha y =: z_\alpha^\tau z_\alpha$，且有以下性质：

1) z_β，z_γ，z_e，z_α 相互独立；

2)　$z_\alpha^\tau z_\alpha \sim \lambda_\alpha^2 \chi_{r_\alpha}^2$，这里 $r_\alpha = a-1$，$\lambda_\alpha^2 = bc\sigma_\alpha^2 + c\sigma_\gamma^2 + \sigma_e^2$，

令 SS_α，SS_β，SS_γ，SS_e 和它们的数学期望相等，就得到一组关于 σ_α^2，σ_β^2，σ_γ^2，σ_e^2 的方程组，解此方程组得方差分量 σ_α^2，σ_β^2，σ_γ^2 和 σ_e^2 的 ANOVA 估计分别为

$$\hat\sigma^2_{\alpha A} = \frac{SS_\alpha}{bcr_\alpha} - \frac{SS_\gamma}{bcr_\gamma},\quad \hat\sigma^2_{\beta A} = \frac{SS_\beta}{acr_\beta} - \frac{SS_\gamma}{acr_\gamma},\quad \hat\sigma^2_{\gamma A} = \frac{SS_\gamma}{cr_\gamma} - \frac{SS_e}{cr_e},\quad \hat\sigma^2_{eA} = \frac{SS_e}{r_e},$$

这里 $r_\alpha = a-1$，$r_\beta = b-1$，$r_\gamma = (a-1)(b-1)$，$r_e = ab(c-1)$。根据上述结论，无论模型 (2) 中 α 是随机效应还是固定效应，方差参数 σ_β^2，σ_γ^2 和 σ_e^2 的 ANOVA 估计一样。所以只需要给出 σ_α^2 的修正 ANOVA 估计。对 σ_α^2，考虑估计族

$$\mathcal{L}_\alpha = \Big\{\frac{m_{\alpha_1}}{bc}\Big(\frac{SS_\alpha}{r_\alpha} - \frac{m_{\alpha_2}SS_\gamma}{r_\gamma}\Big)\Big|\, 0 < m_{\alpha_1} < 1,\ m_{\alpha_2} < 1\Big\}.$$

类似于在估计族 $\mathcal{L}_\beta$ 或者 $\mathcal{L}_\gamma$ 选择权重的方法，经过计算得到 $m_{\alpha_1} = r_\alpha/(r_\alpha+2)$，$m_{\alpha_2} = r_\gamma/(r_\gamma+2)$，相应的得到 σ_α^2 的估计 $\hat\sigma^2_{\alpha M_2}$，它具有如下性质。

定理 3　在均方损失意义下，估计 $\hat\sigma^2_{\alpha M_2}$ 优于 ANOVA 估计 $\hat\sigma^2_{\alpha A}$。

参考文献：

[1] LaMotte L R. On non-negative quadratic unbiased estimation of variance components[J]. Journal of the American Statistical Association, 1973, 68: 728-730

[2] Mathew T, Sinha B K, Sutradhar B C. Nonnegative estimation of variance components in unbalanced mixed models with two components[J]. Journal of Multivariate Analysis, 1992, 42: 77-101

[3] 范永辉，王松桂. 两向分类随机效应模型中方差分量的非负估计[J]. 工程数学学报，2007, 24(2): 303-310

[4] Kelly R J, Mathew T. Improved nonnegative estimation of variance components in some mixed modles with unbalanced data[J]. Technometrics, 1994, 36: 171-181

[5] Wang S G, Chow S C. Advanced Linear Models: Theory and Applications[M]. New York: Marcel Dekker Inc, 1994

New Estimation of Variance Components in Two-way Classification Model with Random Effects

XU Wang-li

(School of Statistics, Renmin University of China, Beijing 100872)

Abstract: That two-way classification model with random effects is used widely in practice, one of the most important estimation methods for this model is analysis of variance estimate (ANOVA). Notice that the mean square error (MSE) of the ANOVA is not the smallest, we propose an improved ANOVA based on new estimation family and prove that the MSE of the new estimator is smaller than that of ANOVA. The estimation method is extended to general models, which are used mainly in the field of medicine.

Keywords: two way classification model with random effects; ANVOA estimator; MSE